What Makes a Genome?

Organisation of Complex Genomes

Dr Ben Skinner

Mammalian Molecular Genetics Group • Department of Pathology

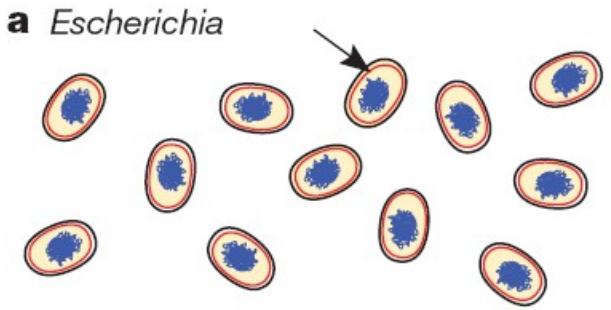
bms41@cam.ac.uk • 01223 333 709

Complex vs simple genomes

Prokaryotic Very diverse genomes!	**Eukaryotic** Also very diverse!
<1Mb to ~10Mb	<10Mb to >100Gb
Linear or circular chromosomes (plus circular or linear plasmids)	Linear chromosomes
No introns in genes; little intergenic content	Introns; can have high amount of intergenic content
Little repetitive content	Can be highly repetitive
Mostly protein coding	Can have very low percentage coding protein
	Mitochondria + chloroplasts with own genome

There is no ‘typical’ prokaryotic or eukaryotic genome!

Origins of complex genomes?



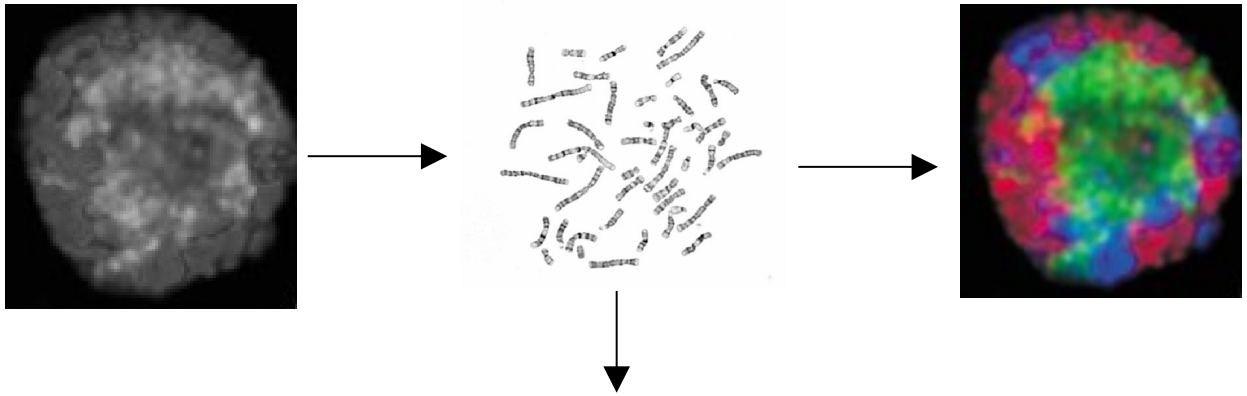
From Nature (2010) ***467****(7318): 929-34*

Aspects of genome organisation

Genome Size

Chromosome Number & Karyotype

Nuclear Organisation



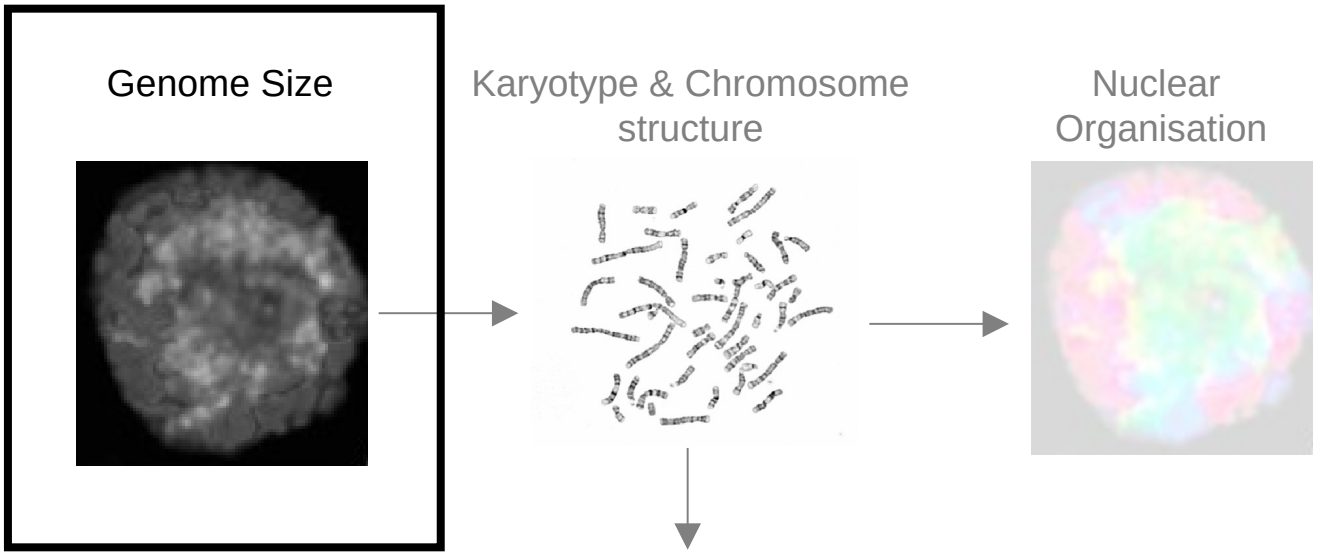
Classes of coding and non-coding DNA

Strutural elements within class

DNA sequence

.....ACGTTTACCGATTAACTAAACGTTTACCCGATTAAGATTAAACTAA.....

Aspects of genome organisation



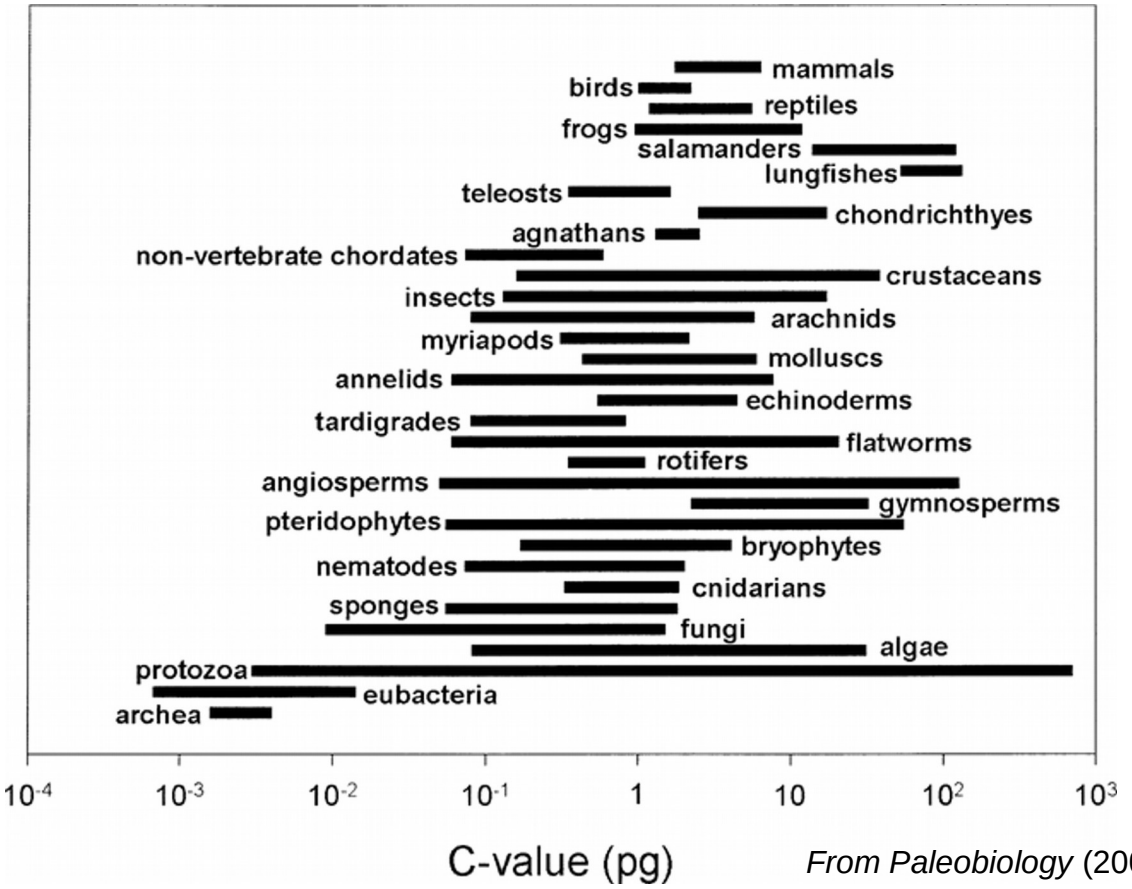
Classes of coding and non-coding DNA

Strutural elements within class

DNA sequence

.....ACGTTTACCGATTAACTAAACGTTTACCCGATTAAGATTAAACTAA.....

Genome size and the C-value enigma

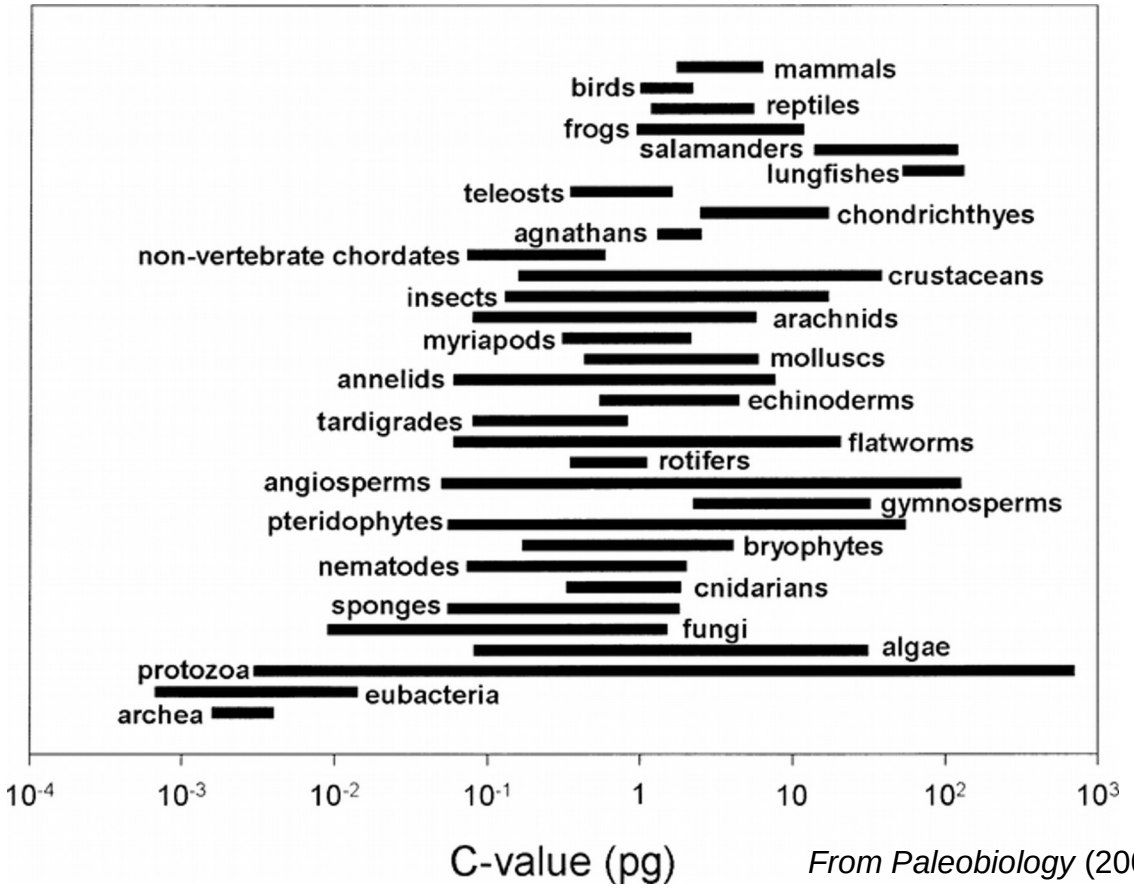


From Paleobiology (2004) ***30****: 179-202*

C-value = the amount of DNA in a haploid nucleus

1pg ~ 978Mb

Genome size and the C-value enigma



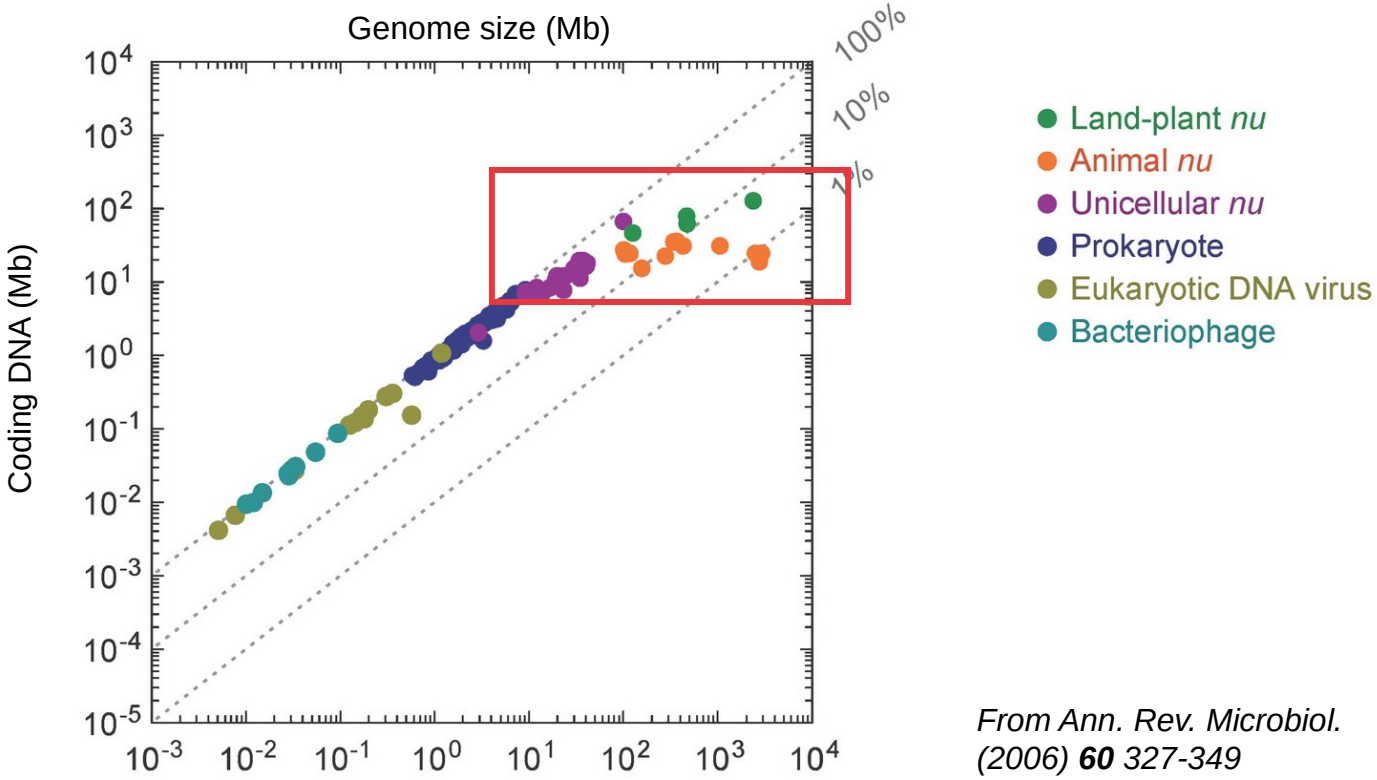
From Paleobiology (2004) ***30****: 179-202*

Genome sizes vary more than 200,000-fold among eukaryotes

Genome size does not correlate with organismal complexity

Genome size and the C-value enigma

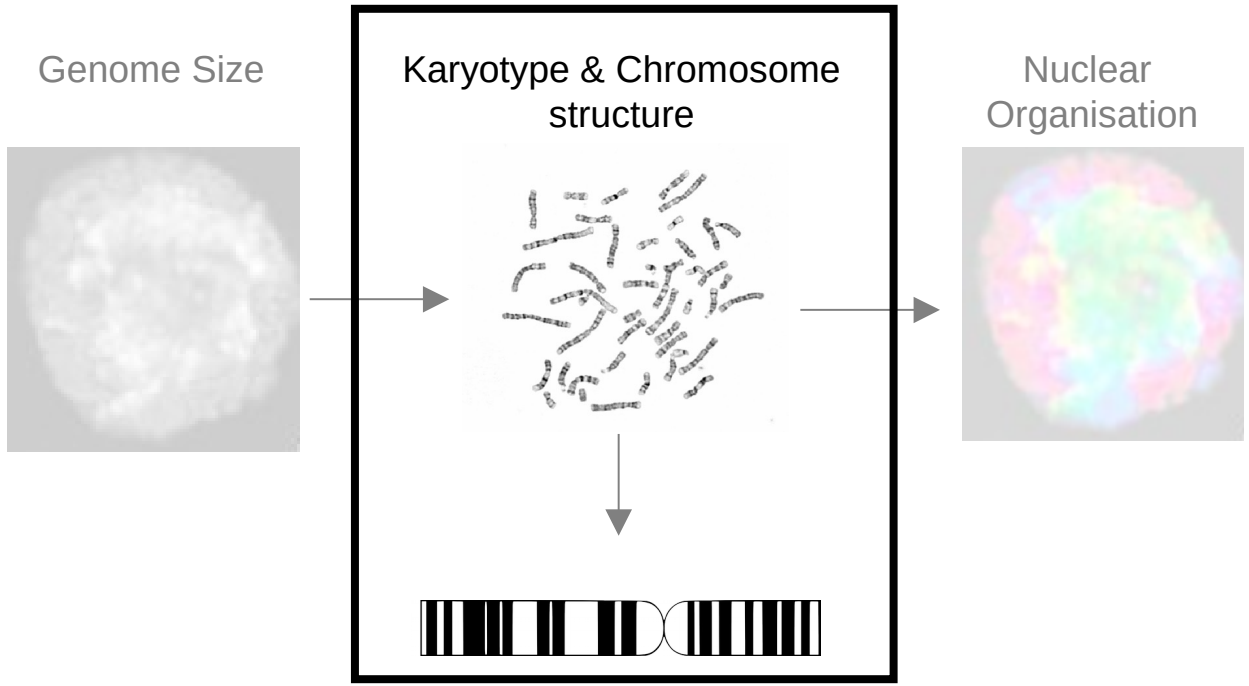
- Genome size does not correlate with organismal complexity in eukaryotes
- Genome size does not correlate with the amount of coding DNA either



From Ann. Rev. Microbiol. (2006) ***60*** *327-349*

Much (or most) of the DNA in eukaryotic cells does not code for proteins

Aspects of genome organisation



Classes of coding and non-coding DNA

Strutural elements within class

DNA sequence

.....ACGTTTACCGATTAACTAAACGTTTACCCGATTAAGATTAAACTAA.....

Chromosome number in eukaryotes

Myrmecia pilosula

Jack jumper ant

2n=1 (♂), 2 (♀)

Haplodiploidy

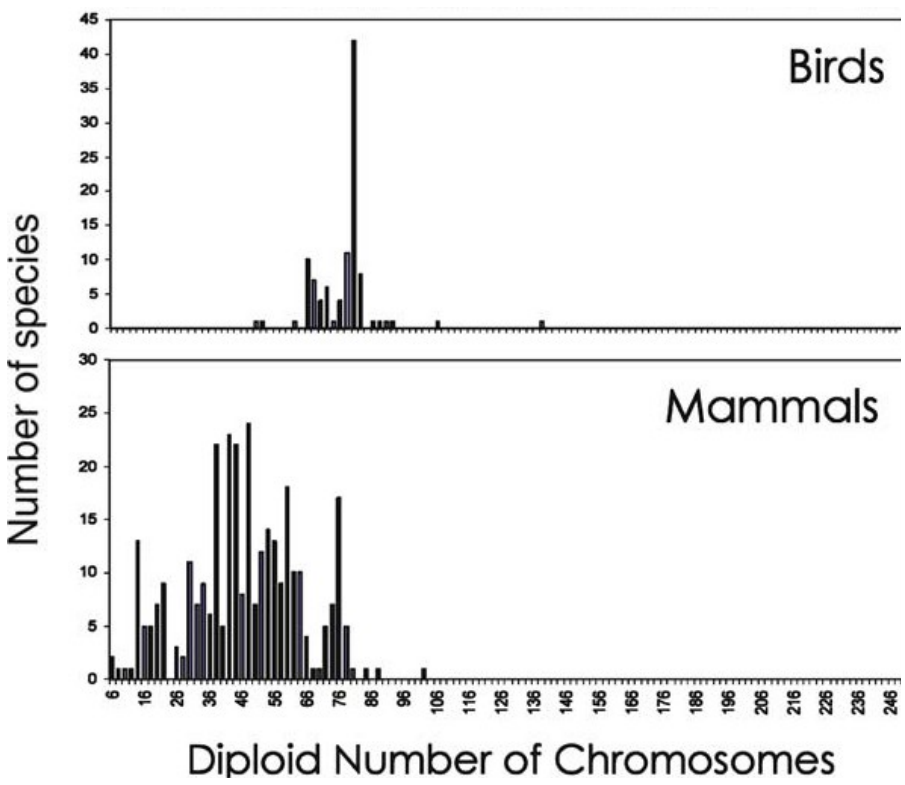
Genus *Ophioglossum*

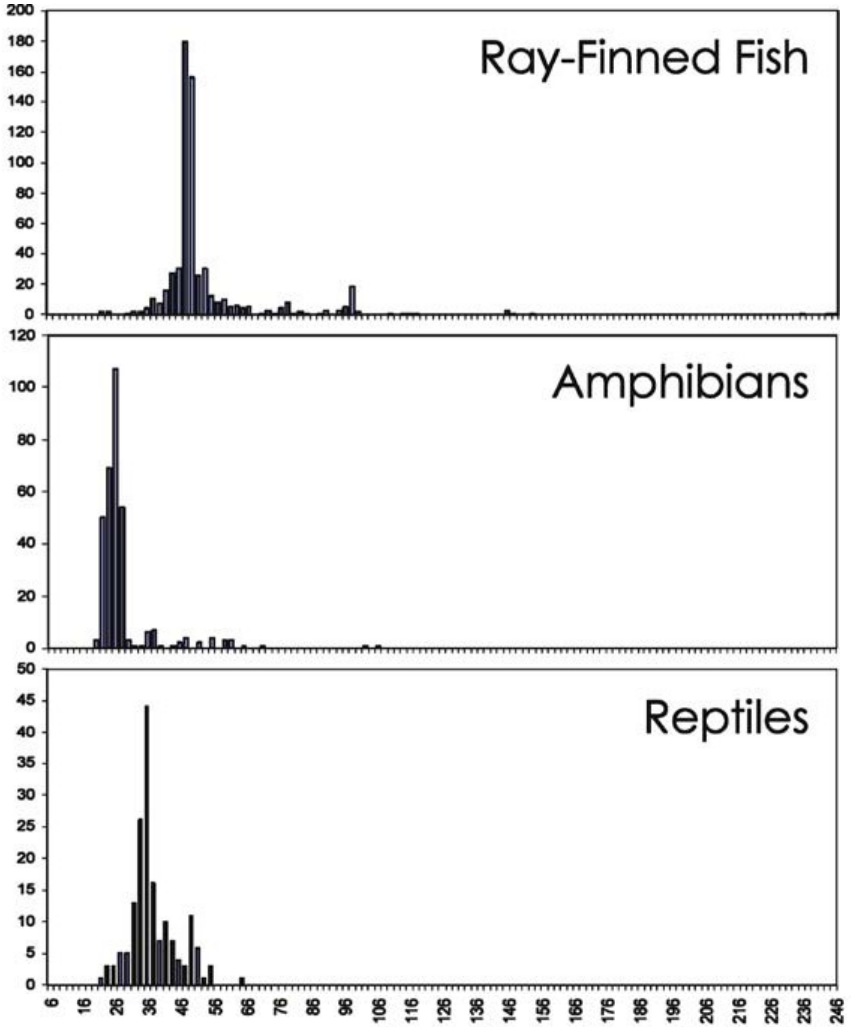
Adder's tongue fern

2n=30-1400

Extreme polyploidy

Chromosome number in vertebrates





*From Genetica (2006) **127**, 321-327*

The organisation of DNA in chromosomes

The diameter of the nucleus of a human cell is about 15 μm

The total length of the naked DNA in this nucleus is about 2 metres

There are about 10 trillion (human) cells in a human body

The total length of DNA in a body is about 20 trillion kilometres

This is 133 times the distance from the earth to the Sun!

How does a cell make this fit?

The organisation of DNA in chromosomes

Chromatin = DNA + proteins

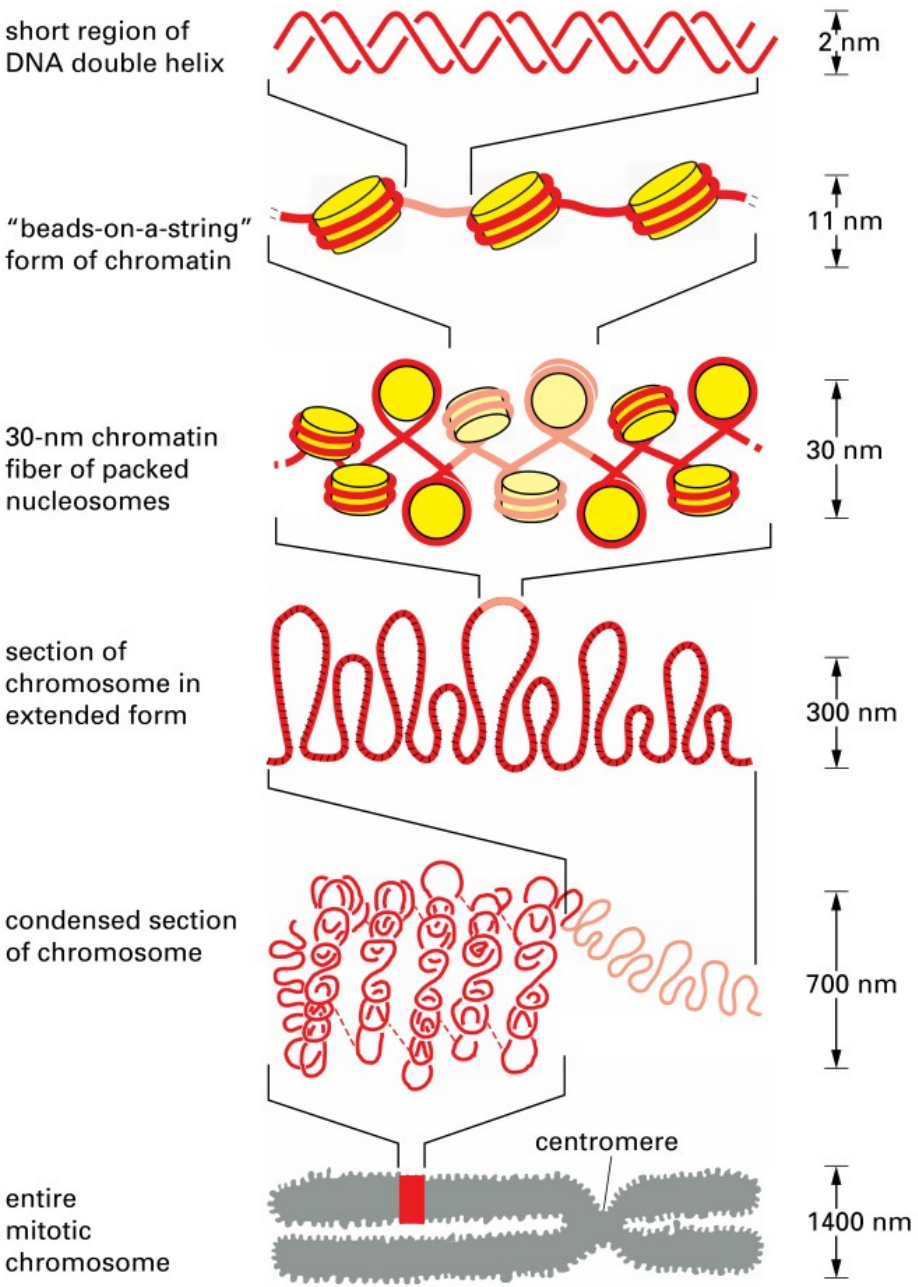
Chromosomal proteins

1) Histones (H1, H2A, H2B, H3, H4)

- small basic proteins with a net positive charge (bind to DNA)
- present in all cell types
- important for chromosome structure
- evolutionarily highly conserved

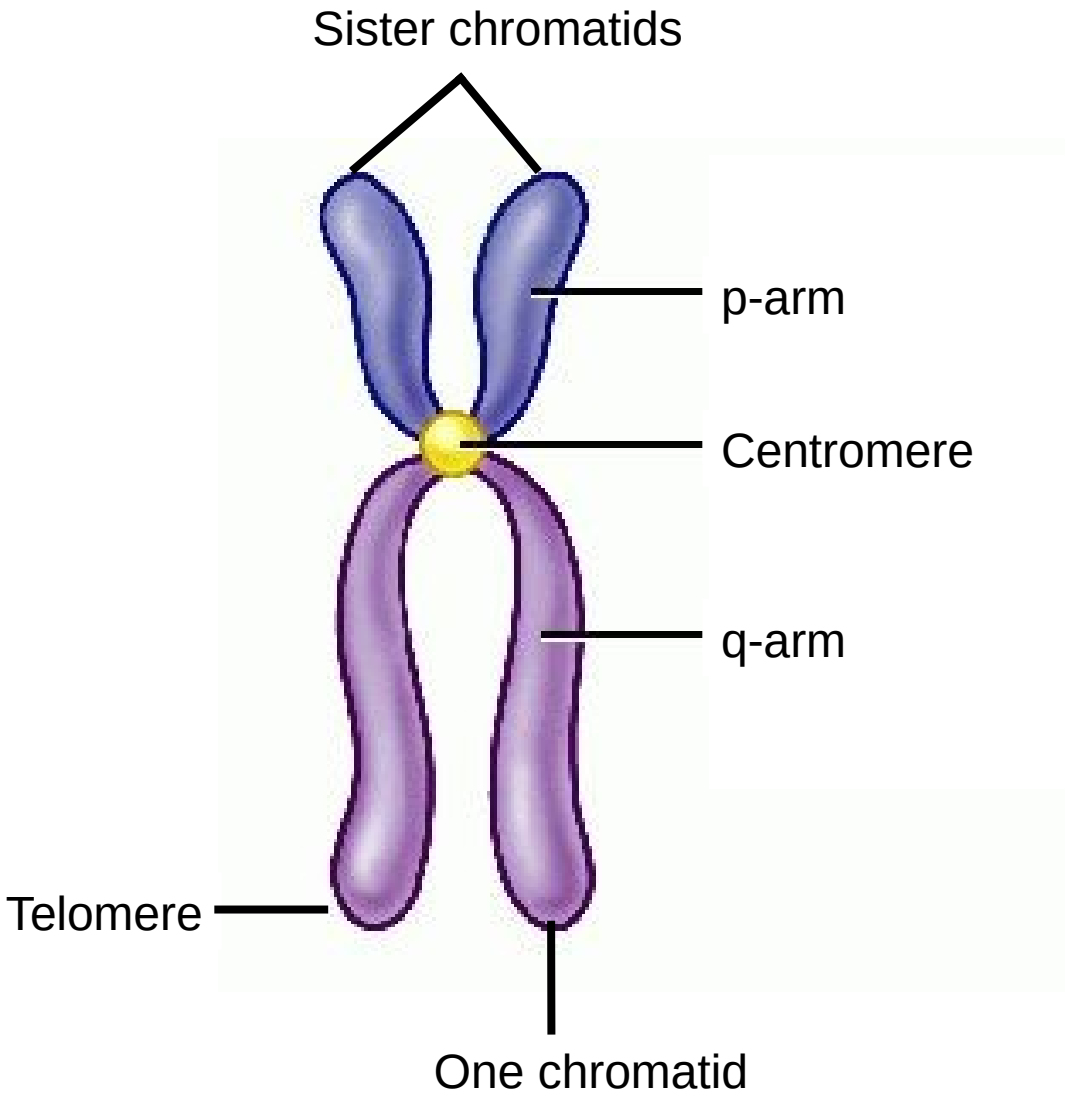
2) Non-histones

- many different types
- often acidic (many bind to histones)
- differ between cell types and organisms



NET RESULT: EACH DNA MOLECULE HAS BEEN PACKAGED INTO A MITOTIC CHROMOSOME THAT IS 10,000-FOLD SHORTER THAN ITS EXTENDED LENGTH

Chromosome structure



Centromeres, telomeres, chromatids

Centromeres

- Kinetochore forms on the centromere
- in mitosis and meiosis, spindle fibres attach to the kinetochore
- DNA sequence of centromeres varies among species

Telomeres

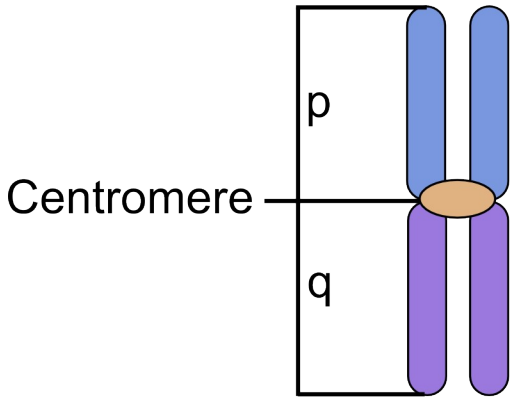
- important for replication and stability of chromosomes
- tandemly repeated simple telomeric sequences (in vertebrates: 5'-TTAGGG-3')

Chromatids

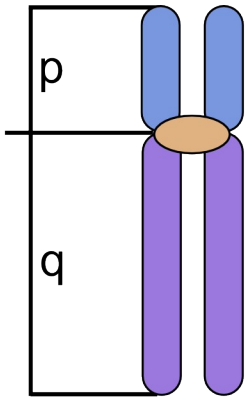
- identical copies
- in mitosis, chromosome splits longitudinally→ one chromatid in each daughter cell

Types of eukaryotic chromosomes

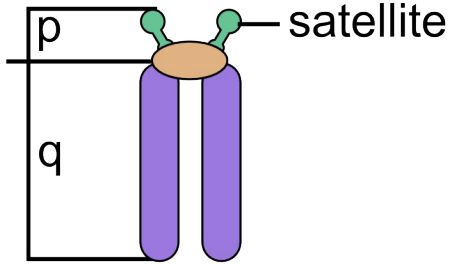
Chromosomes are classified into morphological types according to centromere position:



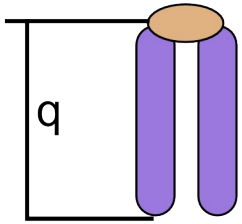
Metacentric



Submetacentric



Acrocentric

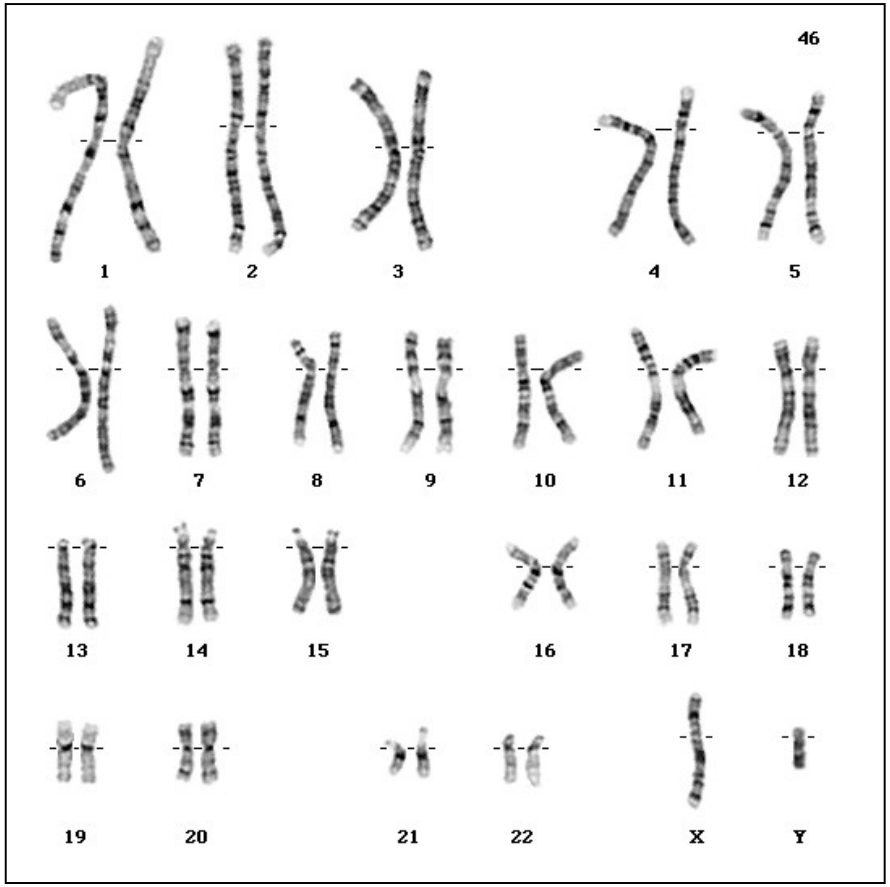


Telocentric

The karyotype

- A karyotype is the complete set of all metaphase chromosomes in a cell
- Humans have 46 chromosomes (23 pairs)

- More in lecture ‘What chromosome studies tell us about disease ’



Autosomes and sex chromosomes

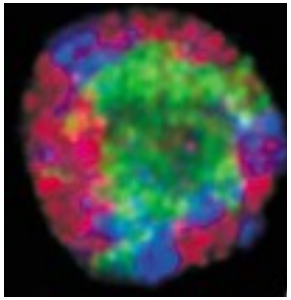
- Many animals and plants have differentiated sex chromosomes
- The chromosomes that are not involved in sex determination are called autosomes
- Mammals including humans have an XY sex chromosome system (males are heterogametic – XY, females are homogametic – XX)
- Birds and some insects have a ZW sex chromosome system (males are homogametic – ZZ, females are heterogametic – ZW)
- Many other systems exist…

Aspects of genome organisation

Genome Size

Karyotype & Chromosome structure

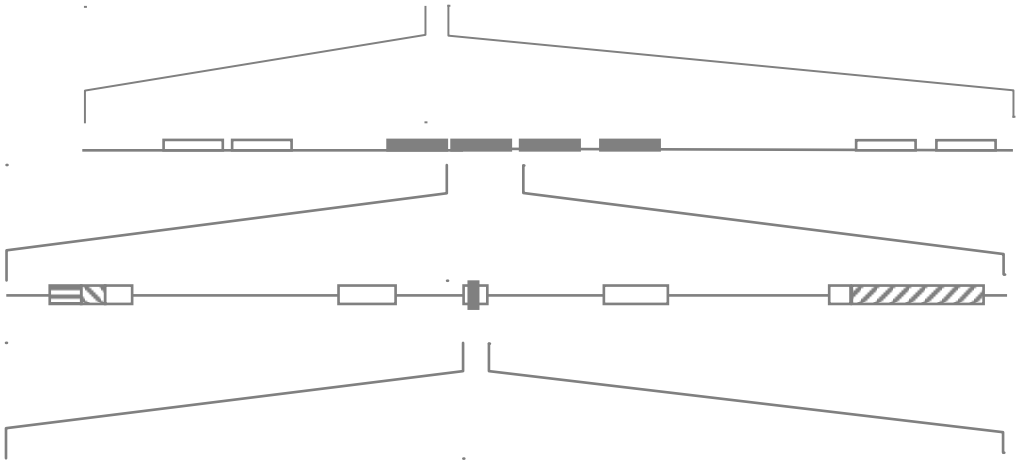
Nuclear Organisation



Classes of coding and non-coding DNA

Strutural elements within class

DNA sequence



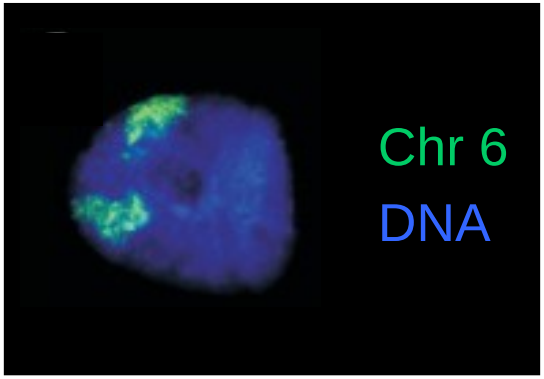
.....ACGTTTACCGATTAACTAAACGTTTACCCGATTAAGATTAAACTAA.....

Nuclear organisation

Positions chromosomes occupy in interphase (non-dividing) nuclei are not random

Organisation can be by:

- Chromosome size
- Gene density
- Some other factor



Implicated in disease, development and evolution

- More in lecture ‘Why chromosome position matters in disease’

Aspects of genome organisation

Genome Size

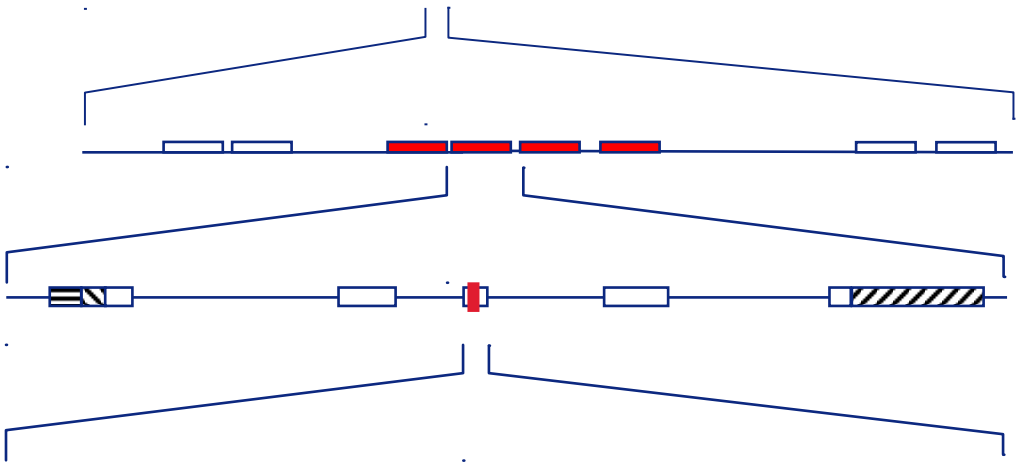
Karyotype & Chromosome structure

Nuclear Organisation

Classes of coding and non-coding DNA

Strutural elements within class

DNA sequence



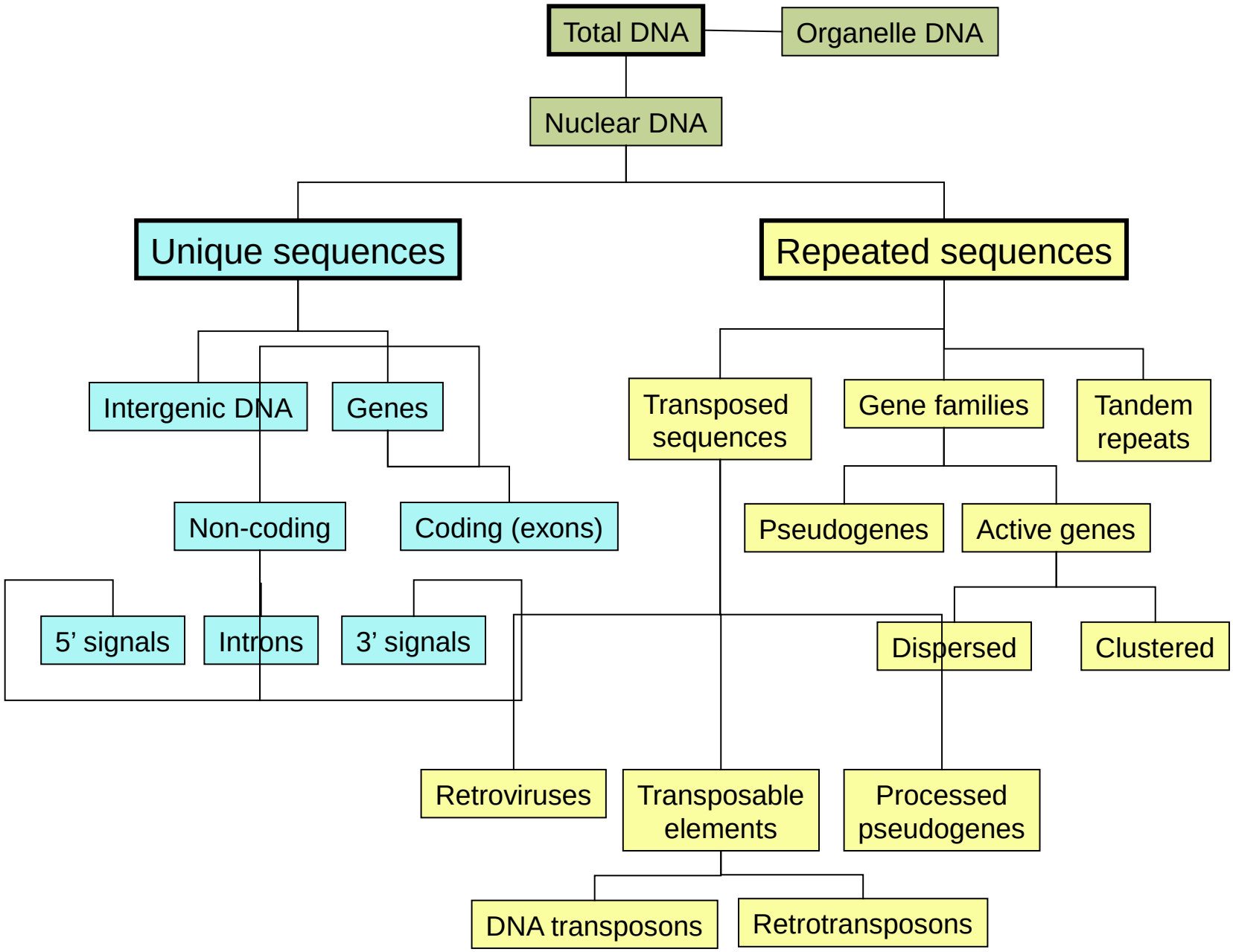
.....ACGTTTACCGATTAACTAAACGTTTACCCGATTAAGATTAAACTAA.....

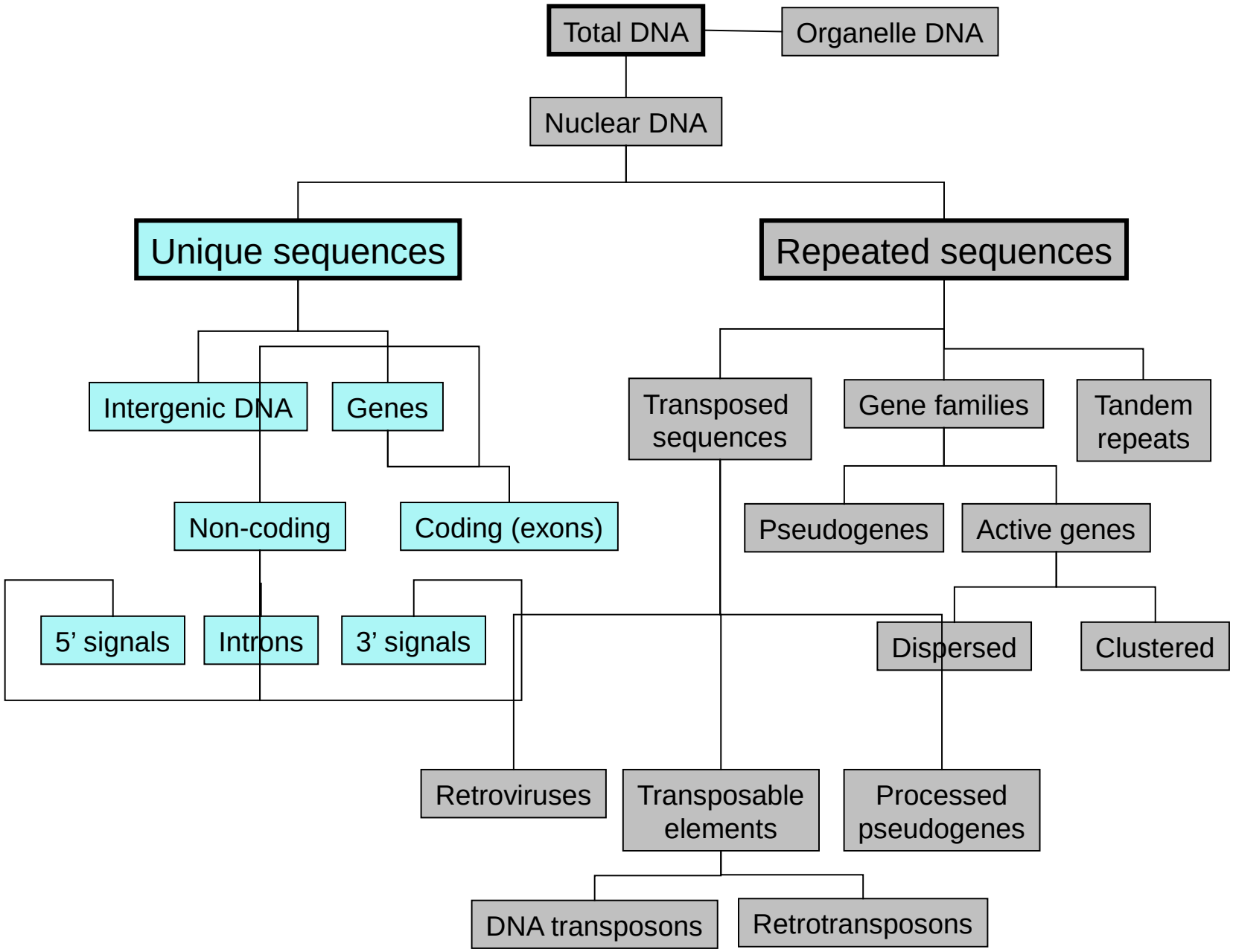
Classes of DNA

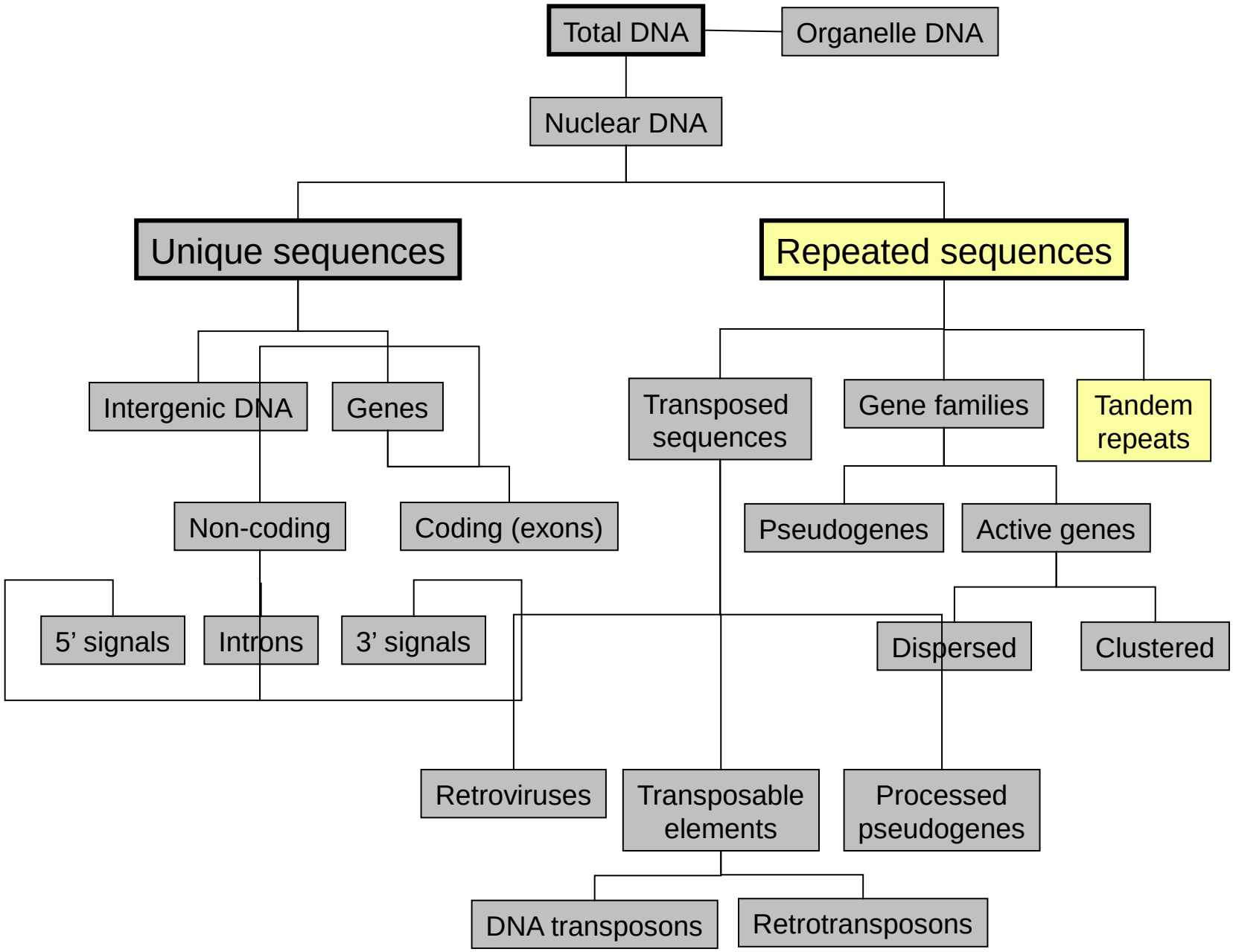
Only ~1.2% of the human genome comprises protein coding genes

~30-35Mb out of 3.1Gb

What is the rest of the genome composed of?


Total DNA
Organelle DNA
Nuclear DNA
Unique sequences
Repeated sequences
Intergenic DNA
Genes
Transposed sequences
Gene families
Tandem repeats
Non-coding
Coding (exons)
Pseudogenes
Active genes
5’ signals
Introns
3’ signals
Dispersed
Clustered
Retroviruses
Transposable elements
Processed pseudogenes
DNA transposons
Retrotransposons


Total DNA
Organelle DNA
Nuclear DNA
Unique sequences
Repeated sequences
Intergenic DNA
Genes
Transposed sequences
Gene families
Tandem repeats
Non-coding
Coding (exons)
Pseudogenes
Active genes
5’ signals
Introns
3’ signals
Dispersed
Clustered
Retroviruses
Transposable elements
Processed pseudogenes
DNA transposons
Retrotransposons


Total DNA
Organelle DNA
Nuclear DNA
Unique sequences
Repeated sequences
Intergenic DNA
Genes
Transposed sequences
Gene families
Tandem repeats
Non-coding
Coding (exons)
Pseudogenes
Active genes
5’ signals
Introns
3’ signals
Dispersed
Clustered
Retroviruses
Transposable elements
Processed pseudogenes
DNA transposons
Retrotransposons

Tandem Repeats

Directly adjacent repeating pattern of 2 or more nucleotides

Minisatellites

- ~10-100bp units
- repeat size can be 0.5-100kb
- found mainly towards telomeric regions in humans

Microsatellites

- ~1-6bp units
- 90% of repeats <40bp total; <2% more than 30 copies
- found throughout the human genome ~25kb intervals

Number of repeats is highly polymorphic between individuals

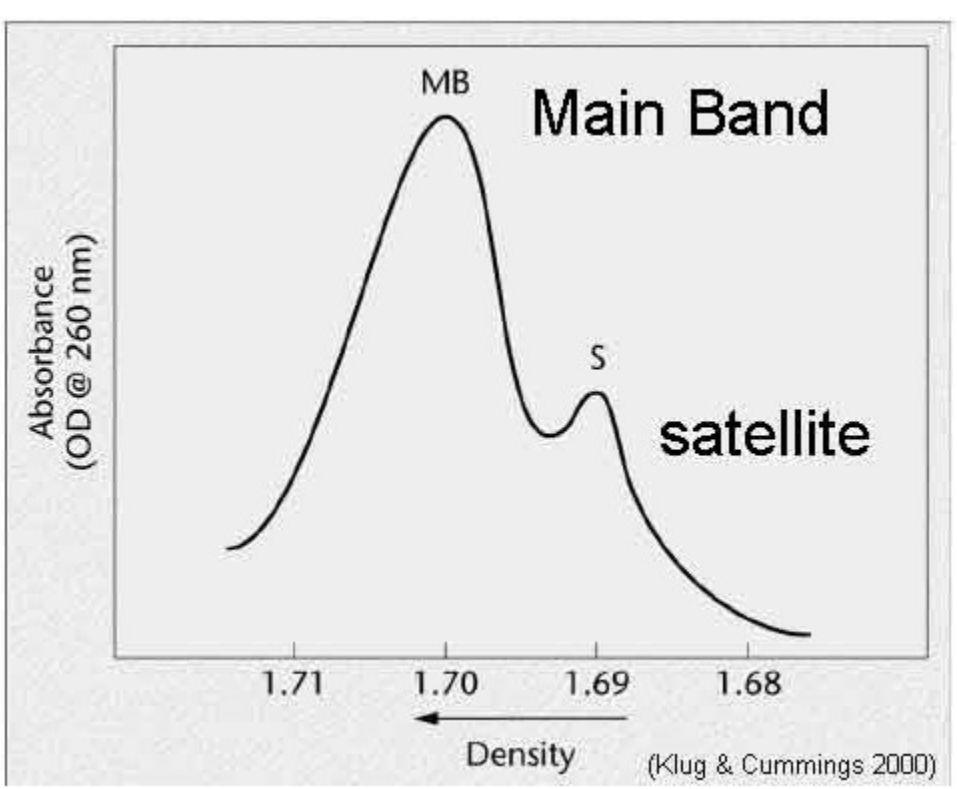
Used for (e.g.):

- DNA fingerprinting
- Forensic analysis
- Paternity testing
- Population genetics

Satellite DNA

- Generally, short, tandemly repeated DNA sequences
  - Includes micro- and mini-satellites

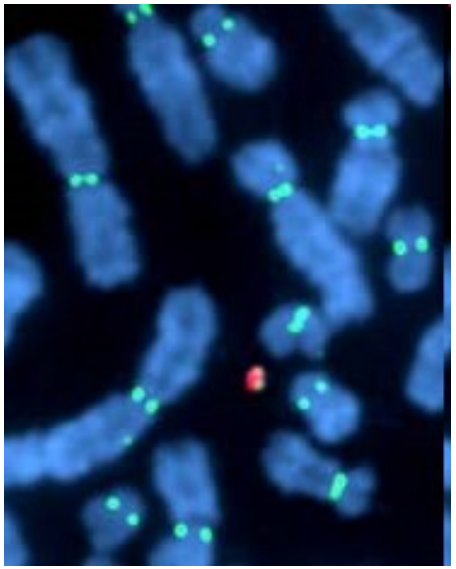
- Named for bands seen when centrifuging sheared DNA in a caesium chloride density gradient



Satellite DNA

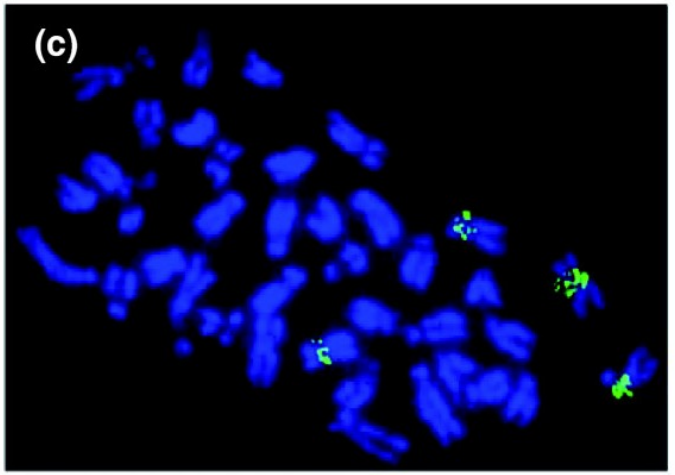
Human satellite DNAs	Unit size (base pairs)	Location
α	171	All chromosomes
β	68	Centromeres of chromosomes 1,9, 13, 14, 15, 21, 22, Y
Satellite 1	25-48	Centromeres & heterochromatic regions
Satellite 2	5	Most chromosomes
Satellite 3	5	Most chromosomes

Satellite DNA



Centromeric α satellite
on human metaphase

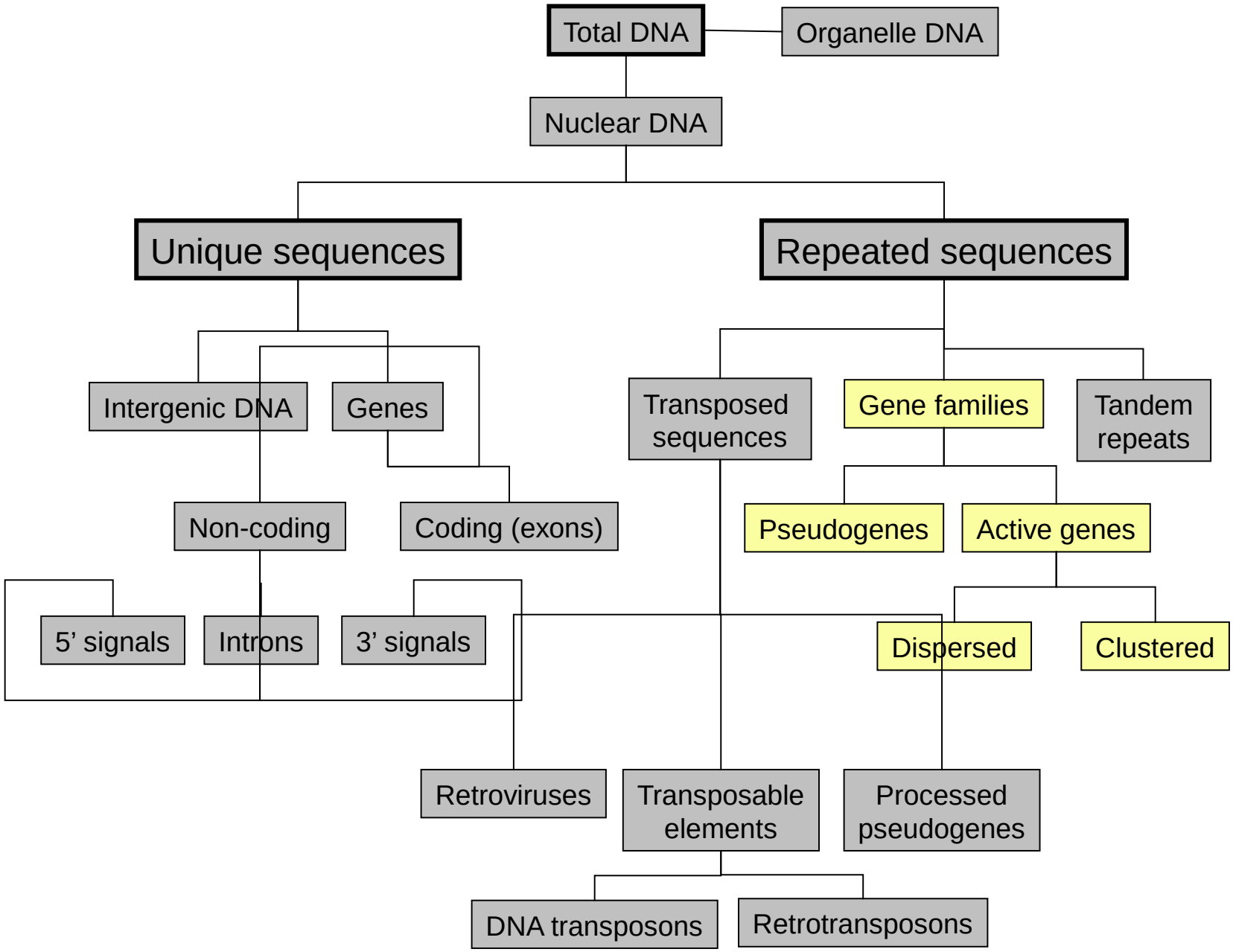
Turquoise killifish, *Nothobranchius furzeri*



Genome Biology (2009) **10** R16

A GC-rich, 24-nucleotide minisatellite
specifically staining centromeric
regions of two chromosome pairs

Total sizes of satellite repeats can be several Mb


Total DNA
Organelle DNA
Nuclear DNA
Unique sequences
Repeated sequences
Intergenic DNA
Genes
Transposed sequences
Gene families
Tandem repeats
Non-coding
Coding (exons)
Pseudogenes
Active genes
5’ signals
Introns
3’ signals
Dispersed
Clustered
Retroviruses
Transposable elements
Processed pseudogenes
DNA transposons
Retrotransposons

Pseudogenes (Ψ)

- Genes rendered inactive, for example by a frameshift mutation, nonsense mutation etc.

- May be inactivation of a single copy gene
  - loss of functionality
  - *e.g. human caspase-12*

- May be inactivation of one copy of a duplicated gene (non-processed pseudogene)
  - one copy retains function
  - second copy accumulates mutations
  - *e.g. multiple copies of human FRG1*

Active gene families

- Genes resulting from individual gene duplication or whole genome duplication
  - paralogs: homologous genes separated by gene duplication
  - orthologs: homologous genes separated by speciation

Homologous = sharing a common ancestor

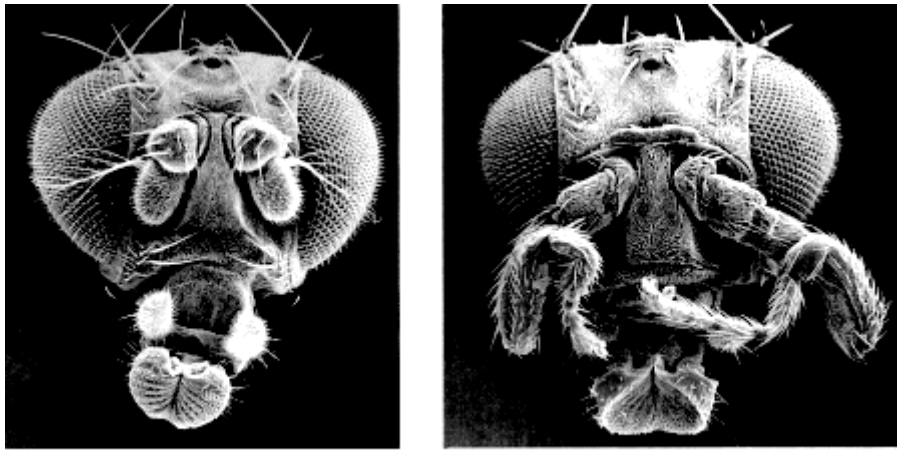
- Copies can take on new functionality / subfunctionality

- Copies may be dispersed throughout the genome
  - *e.g. human ribosomal gene clusters*
    - *chromosomes 13, 14, 15, 21, 22*
    - *~150-200 copies of 14Kb unit (2-3Mb total)*

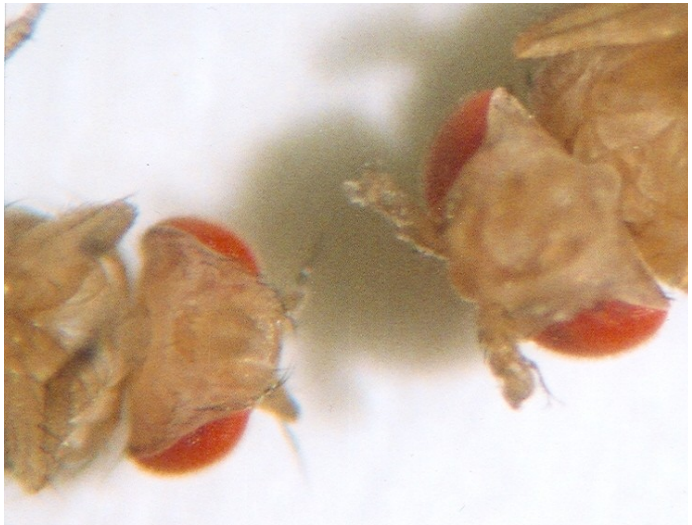
- Or clustered in one region
  - *e.g. invertebrate Hox gene clusters*

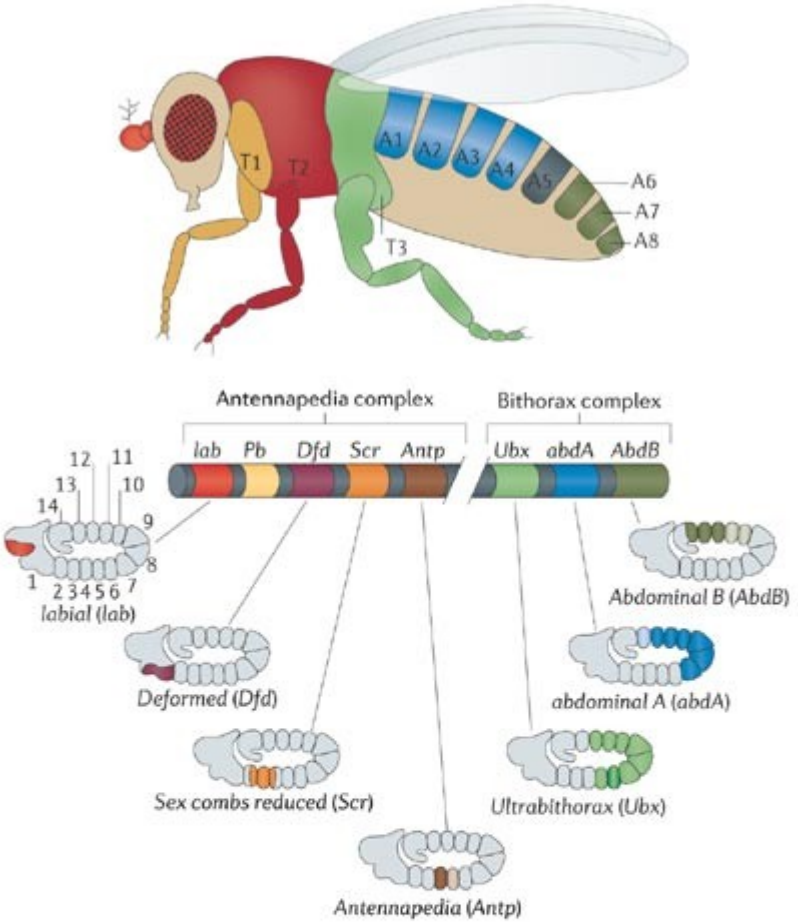
Gene / genome duplication is a mechanism for dispersal of clustered gene families

Hox gene clusters

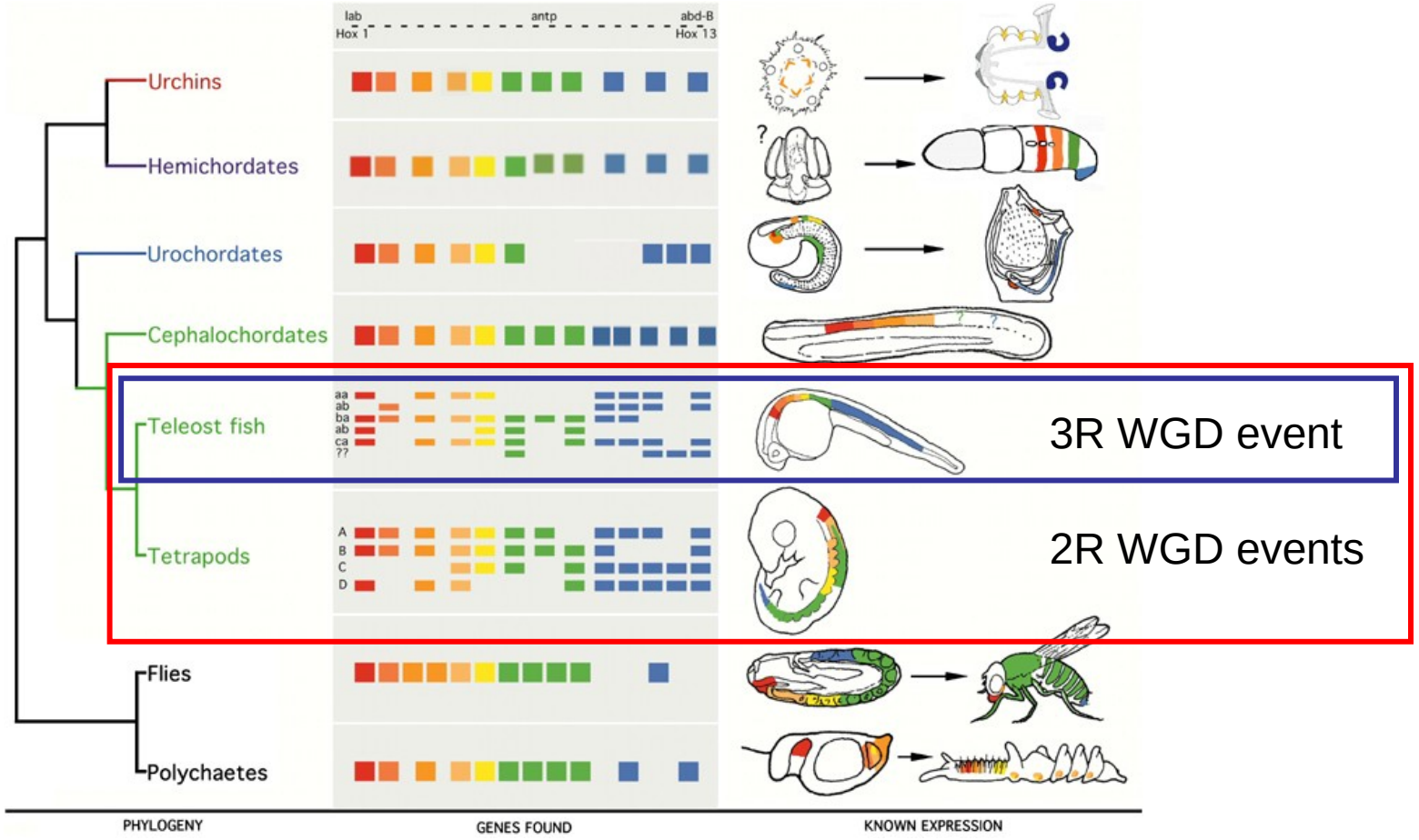


Drosophila: wildtype on left. Right is antennapedia mutant with fully developed legs in place of antennae. Photo by FR Turner, Indiana Univ.

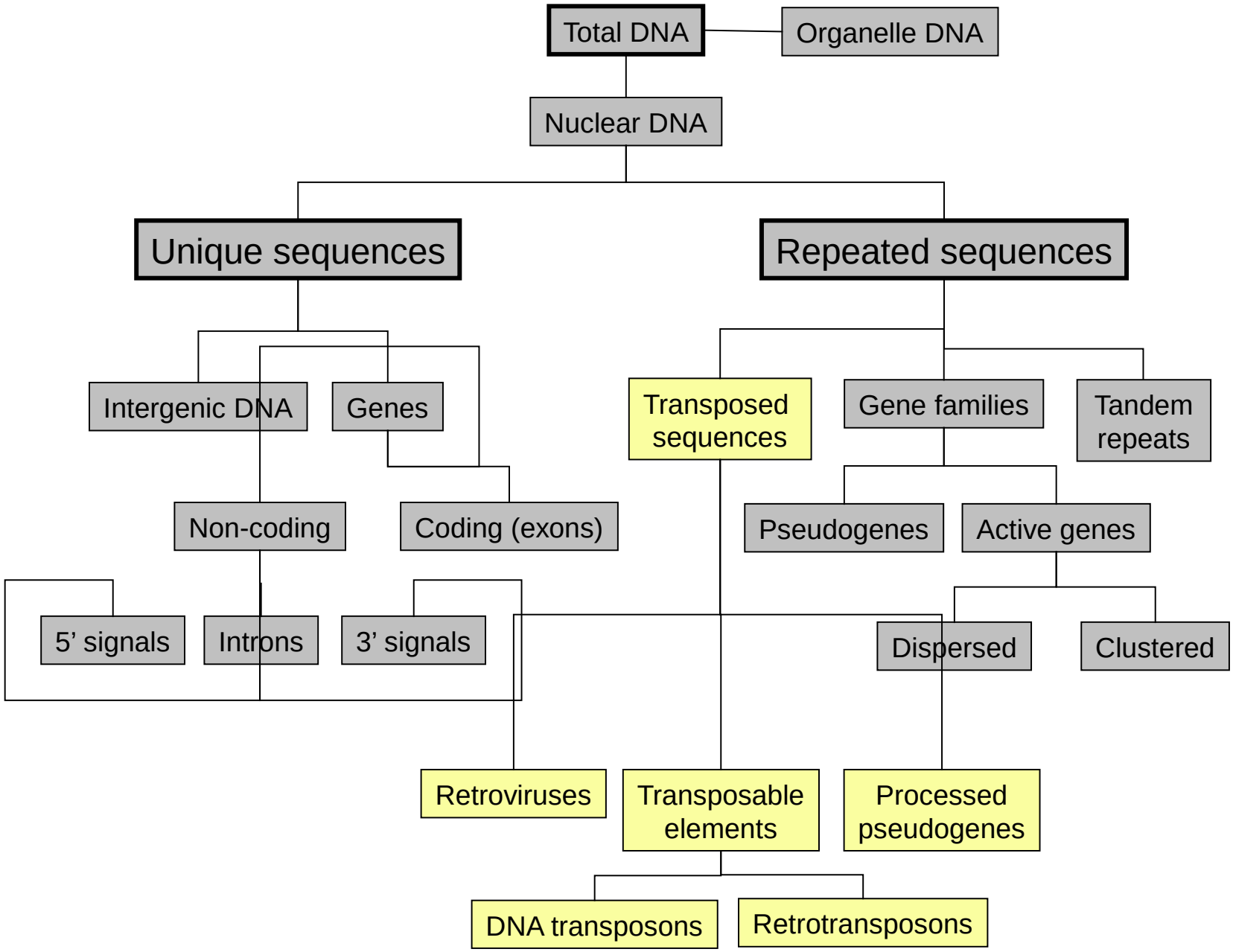




Hox gene clusters: Whole genome duplication



From Heredity (2006) ***97,*** *235–243*


Total DNA
Organelle DNA
Nuclear DNA
Unique sequences
Repeated sequences
Intergenic DNA
Genes
Transposed sequences
Gene families
Tandem repeats
Non-coding
Coding (exons)
Pseudogenes
Active genes
5’ signals
Introns
3’ signals
Dispersed
Clustered
Retroviruses
Transposable elements
Processed pseudogenes
DNA transposons
Retrotransposons

Transposed sequences

Retroviruses

- found in all mammalian genomes
- endogenous retroviruses (ERVs) ~5-8% of human genomes

Transposable elements

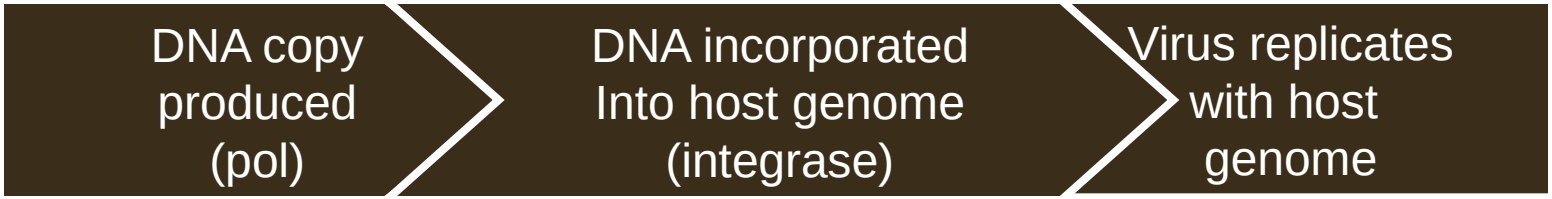
- pieces of DNA which can move around the genome
  - Class I: Retrotransposons – ‘copy and paste’ via RNA intermediate
  - Class II: DNA transposons – often ‘cut and paste’

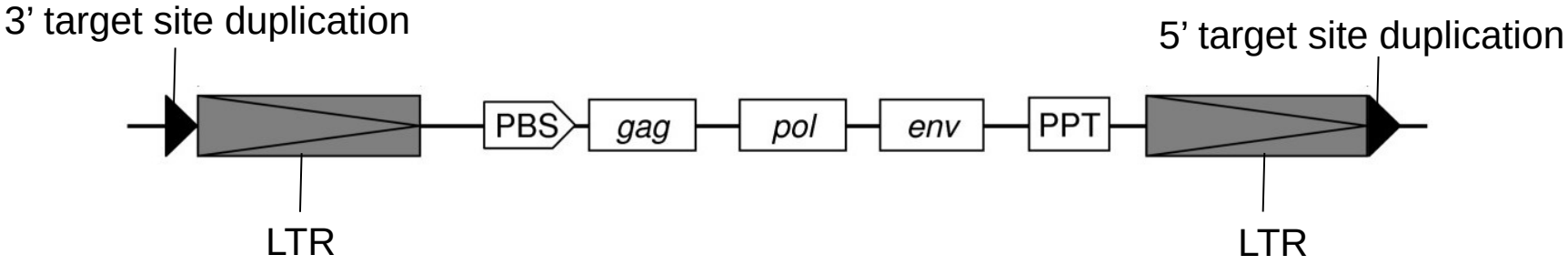
Processed pseudogenes

- integration of cDNA back into genome
- have polyA tail, lack introns and promoter
- occur when random genes get caught in retrotransposition

Retroviruses

RNA genome

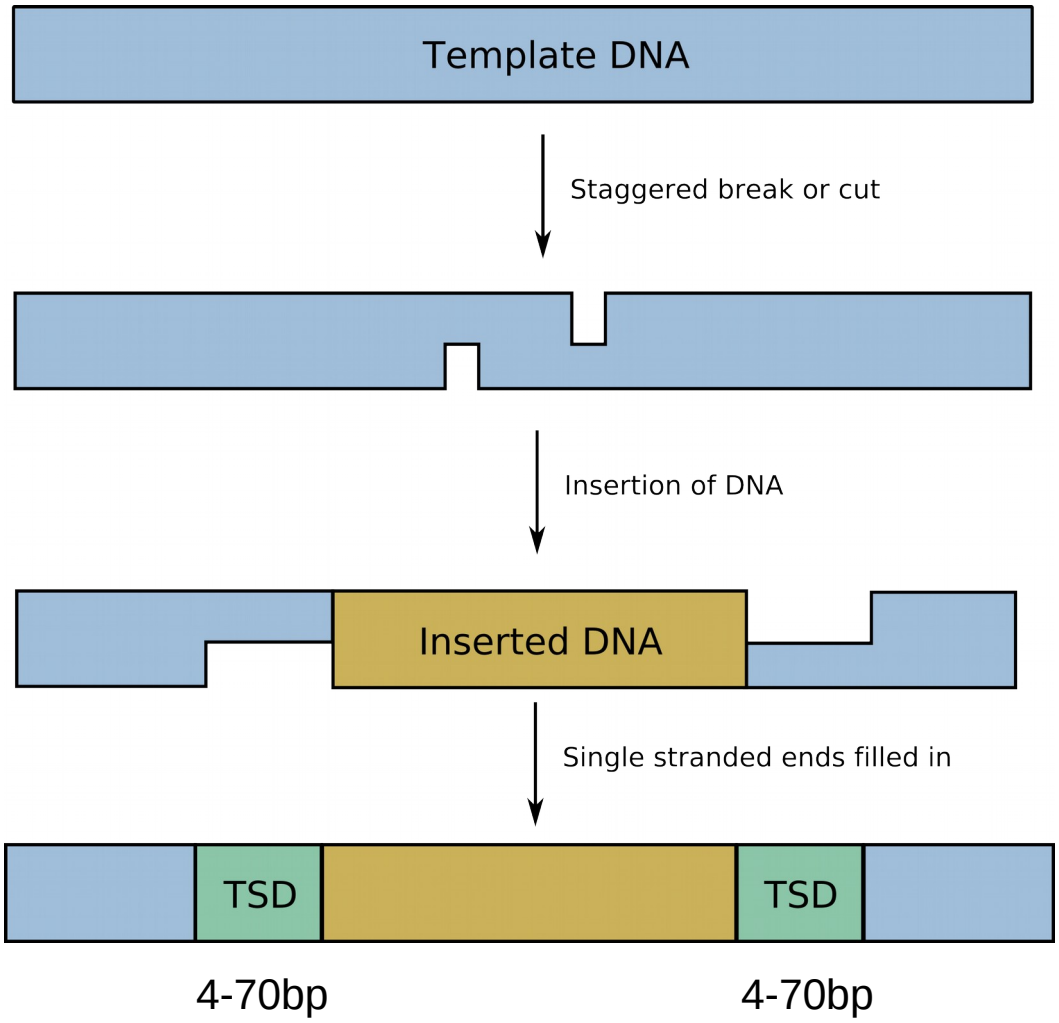




*From BMC Bioinformatics (2008) **9**, 18*

PBS	primer binding site (tRNA binds to initiate reverse transcription)
gag	group specific antigen (DNA binding)
pol	reverse transcriptase
env	envelope protein
PPT	polypurine tract (facilitates reverse transcriptase binding)

Target site duplication (TSD)



Retrotransposons

Class I TEs

Transcription to RNA intermediate → Reverse transcriptase produces dsDNA → Integration into host genome

Two types:

LTR (Long Teminal Repeat) – Virus like

- similar to retroviruses
- do not usually form infectious particles (lack *env*)

From Nature Reviews Genetics (2007) 8, 272-285

Non-LTR

- LINEs (Long Interspersed Nuclear Elements)
- SINEs (Short Interspersed Nuclear Elements)

LINEs and SINEs

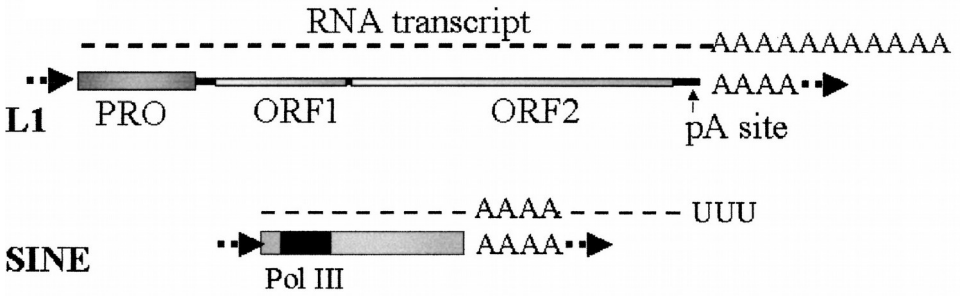
Non-LTR retrotransposons

LINEs (Long Interspersed Nuclear Elements)

- encode reverse transcriptase and often endonuclease
- ~21% of human genome (predominantly non-functional)
- *e.g. L1*

SINEs (Short Interspersed Nuclear Elements)

- <500bp
- do not encode functional protein; rely on other mobile elements for transposition
- ~13% of human genome
- *e.g. Alu*



From Genome Res (2002) 12(10), 1455–1465

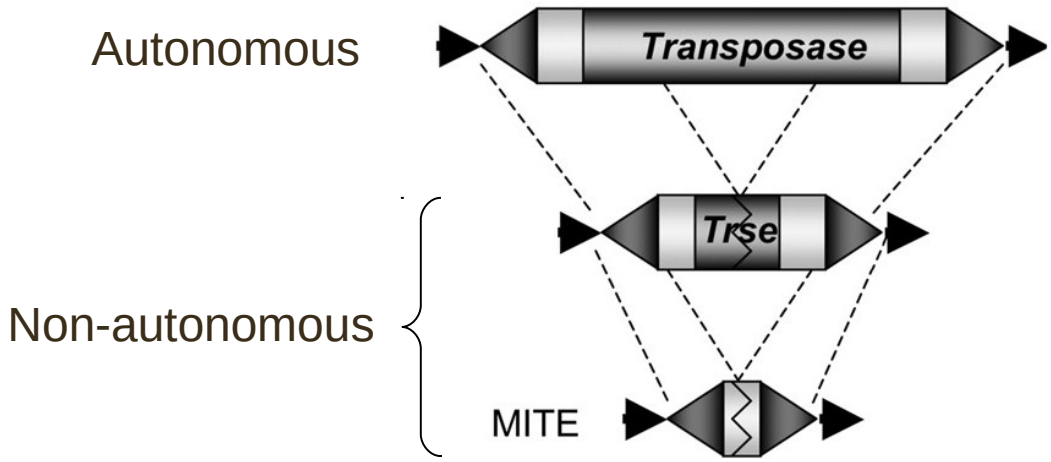
DNA transposons

Class II TEs

DNA sequences that move around the genome

- Transposition (different to retro-transposition in that there is no RNA intermediate)
- Usually ‘cut and paste’ though can be replicative

Encode a transposase enzyme:



*From PNAS (2006)**103** (47) 17600-17601*

Processed pseudogenes

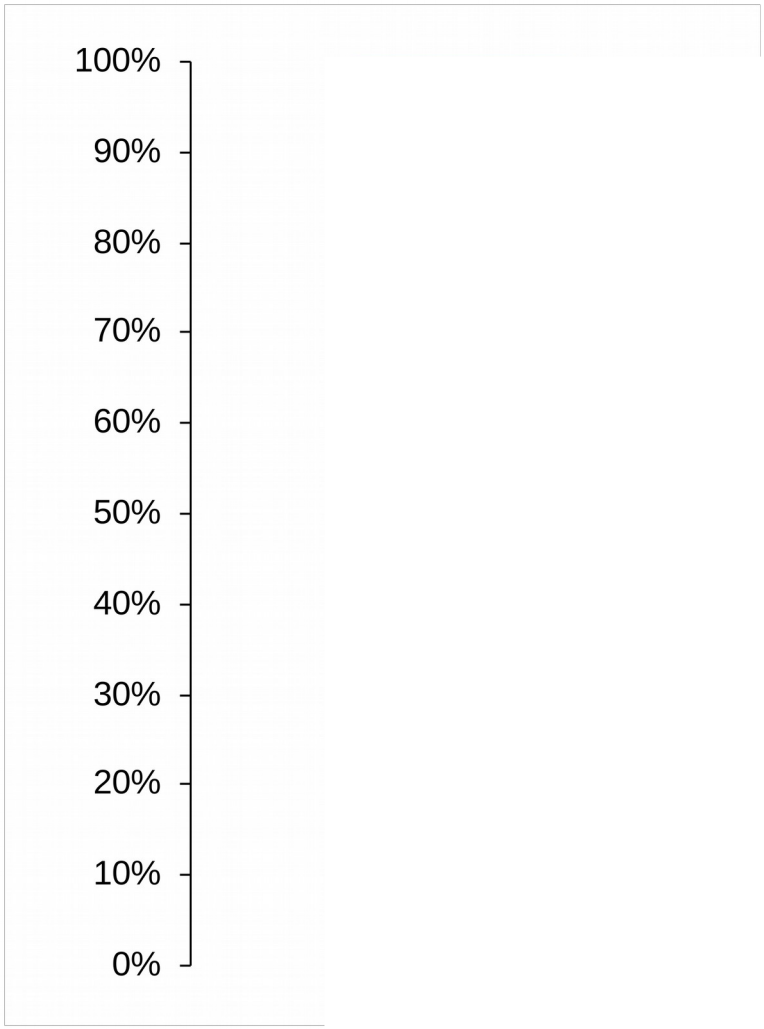
- Retrotransposition usually creates copies of the transposable element
- Occasionally another mature mRNA is reverse-transcribed and integrated instead
- This inserted copy lacks a promoter
  - therefore ‘dead on arrival’
- Commonly have a 5’ truncation due to low processivity of the reverse transcriptase

Processivity = average number of nucleotides added by a polymerase per association/dissociation with the template

Processed pseudogenes

Non-processed pseudogenes	Processed pseudogenes
Usually near active copy	Dispersed
No TSD	TSD
Introns	No introns
No poly A tail	Poly A tail
May have promoter	No promoter
Usually no 5’ deletion	5’ deletion
Arise via gene duplication	Arise via RNA intermediate

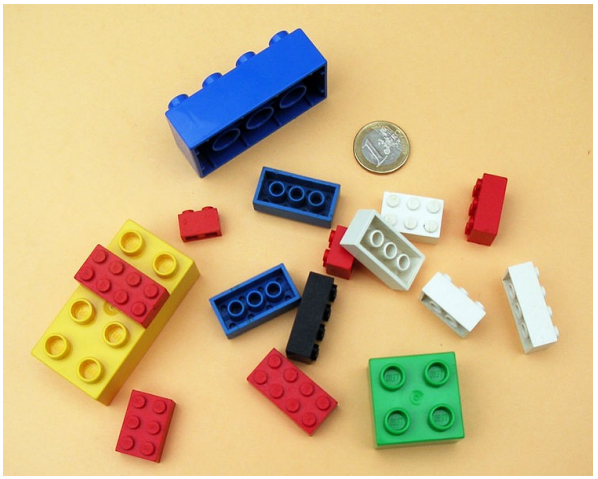
So what are our genomes made of?



Potential structural roles not dependent on sequence identity?

Junk DNA

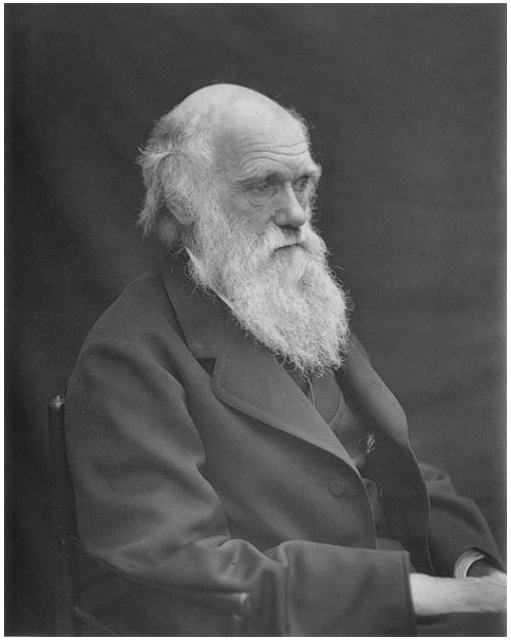
vs



No selection pressure to eliminate useless DNA unless actively detrimental

Junk DNA – The Onion Test

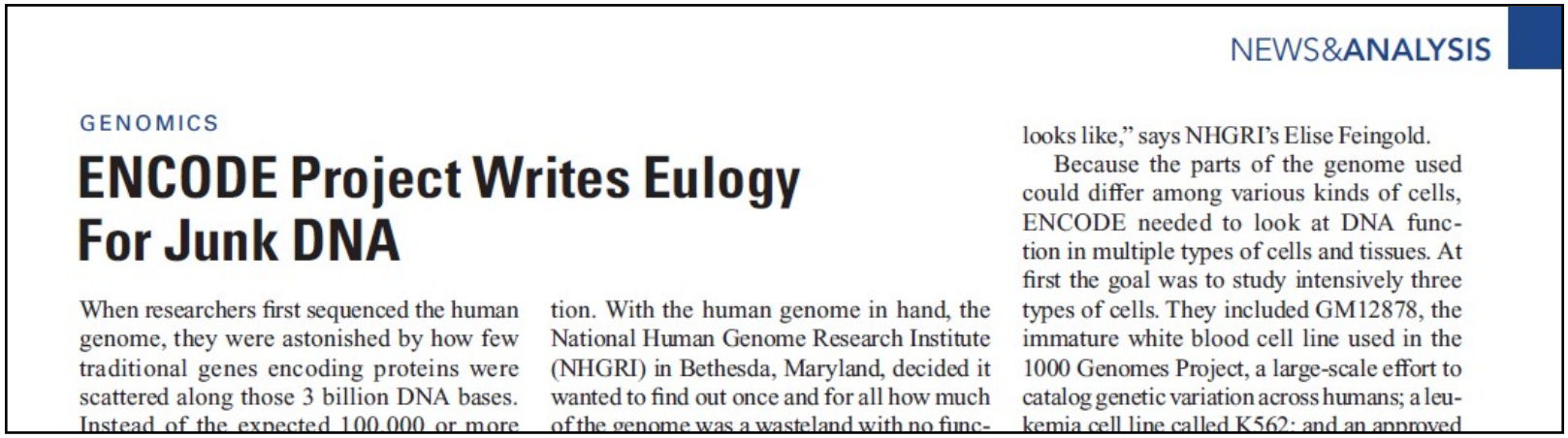
Haploid genome size
~17pg



Haploid genome size
~3.5pg

If all non-coding DNA is functional, why does an onion need 5x as much DNA for this function as a human?

Junk DNA & ENCODE



NEWS&ANALYSIS

GENOMICS

ENCODE Project Writes Eulogy For Junk DNA

When researchers first sequenced the human genome, they were astonished by how few traditional genes encoding proteins were scattered along those 3 billion DNA bases. Instead of the expected 100,000 or more

tion. With the human genome in hand, the National Human Genome Research Institute (NHGRI) in Bethesda, Maryland, decided it wanted to find out once and for all how much of the genome was a wasteland with no func-

looks like," says NHGRI's Elise Feingold.

Because the parts of the genome used could differ among various kinds of cells, ENCODE needed to look at DNA function in multiple types of cells and tissues. At first the goal was to study intensively three types of cells. They included GM12878, the immature white blood cell line used in the 1000 Genomes Project, a large-scale effort to catalog genetic variation across humans; a leukemia cell line called K562; and an approved

ENCODE project recently published a comprehensive analysis of biochemical activity across human genomes

Many media sources reported this as the end of the concept of junk DNA

However: broad definition of activity – e.g. binds transcription factors - does not imply biological function! A more realistic value is 20% of biologically relevant DNA

What can be confidently said to be biologically functional: a *"conservative estimate of our expected coverage of exons + specific DNA:protein contacts gives us 18%, easily further justified (given our sampling) to 20%"*

– Ewan Birney, ENCODE project co-ordinator

Gene death and rebirth: *IRGM*

- Immunity Related GTPase gene family

- 3 copies of IRG gene family in most mammals
  - The families are comprised of gene clusters (e.g. *IRGM* cluster, *IRGC* cluster)
  - Humans however have only 2 genes (*IRGC* & truncated *IRGM*)

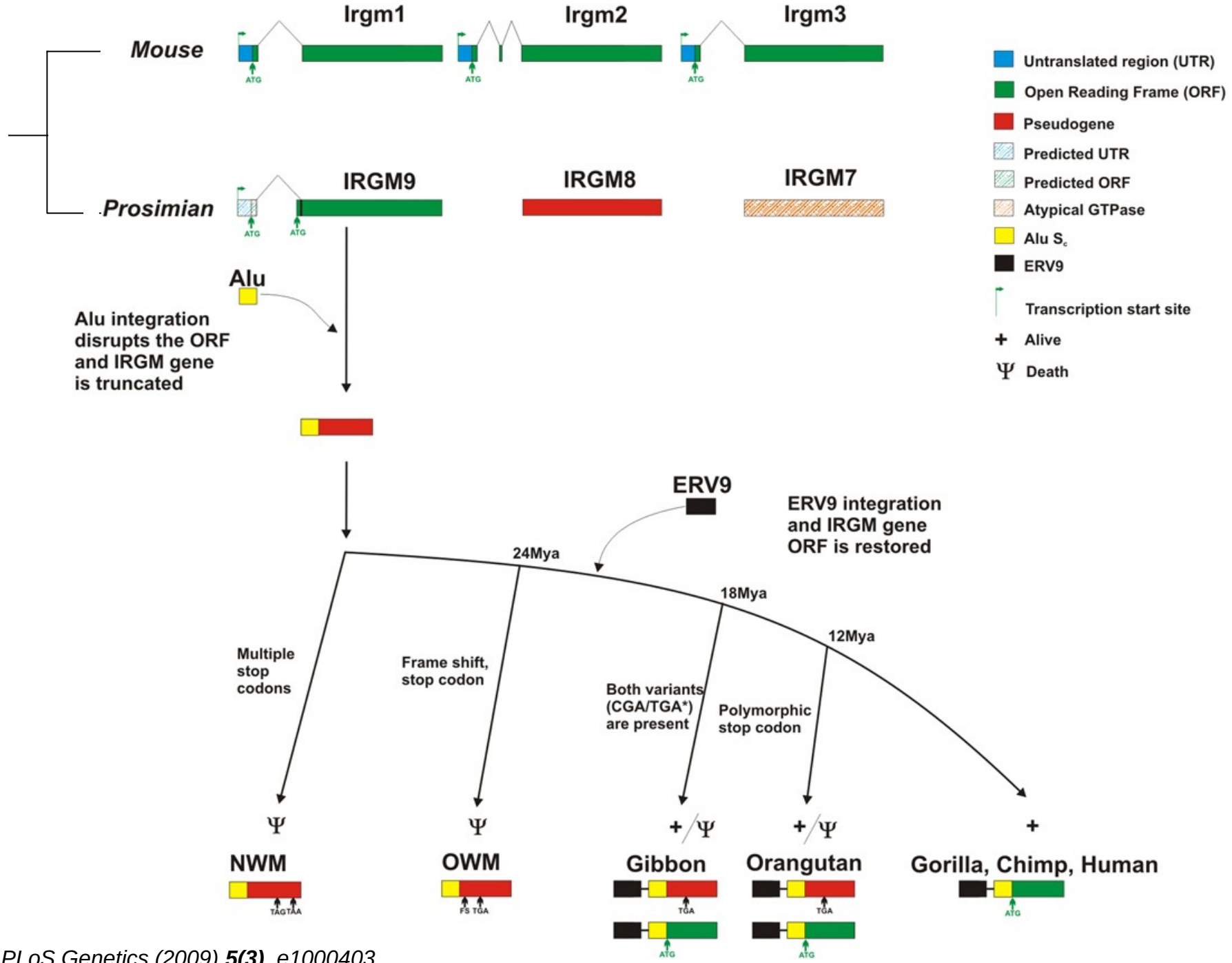
50Mya
- IRGM cluster: inactivation of all but one copy in monkey + great ape ancestor
  - Insertion of Alu renders *IRGM* a pseudogene via frameshift mutation

24Mya
- ERV9 inserts at start of gene and forms new promoter in great ape ancestor
  - Open reading frame restored

12Mya
- Functional copy fixed in gorilla, chimp and human lineage

Today
- Expressed in several tissues in humans inc. testis
- Some haplotypes associate with increased risk for Crohn's disease

PLoS Genetics (2009) ***5(3)****, e1000403*



PLoS Genetics (2009) ***5(3)****, e1000403*

Summary

- Eukaryotic genomes have multiple levels of organisation

- Studies of genomes must consider all these levels for a coherent understanding of normal phenotypic variation, disease states and evolutionary processes
    - What we see today (e.g. *IRGM*) has arisen from interactions of these features over time

- Very little of a complex genome is coding DNA; much is comprised of selfish genetic elements, regulatory elements and other features
    - These can all influence genome regulation

- The human genome is not the same as all eukaryotic genomes!
    - Each organism’s genome (within and between species) has unique features

Further reading

Chromosome evolution in eukaryotes: a multi-kingdom perspective
Coghlan et al. (2005) *Trends in Genetics* 21(12) 673 - 682

Complex human chromosomal and genomic rearrangements
Zhang et al. (2009) *Trends in Genetics* 25(7) 298 - 307

The impact of retrotransposons on human genome evolution
Cordaux & Batzer (2009) *Nature Reviews Genetics* 10(10) 691-703

L1 elements, processed pseudogenes and retrogenes in mammalian genomes
Ding et al. (2006) *IUBMB Life* 58(12) 677-85

The life and death of gene families
Demuth & Hahn (2009) *Bioessays* 31(1) 29-39

Contrasting evolutionary dynamics between angiosperm and mammalian genomes.
Kejnovsky et al. (2009) Trends in Ecology and Evolution 24(10) 572-82

Any recent genetics textbook will also cover the material in this lecture e.g.:

Genes IX, Lewin – a detailed gene-centric textbook
Genomes 3, Brown – broader overview of genomes and genomics